US 20030172043A1

(19) United States

(12) Patent Application Publication

(10) Pub. No.: US 2003/0172043 A1

Guyon et al.

(43) Pub. Date: Sep. 11, 2003

(54) **METHODS OF IDENTIFYING PATTERNS IN BIOLOGICAL SYSTEMS AND USES THEREOF**

(76) Inventors: **Isabelle Guyon**, Berkeley, CA (US); **Jason Weston**, St. Leonard's on Sea (GB)

Correspondence Address:
JOHN S. PRATT, ESQ
KILPATRICK STOCKTON, LLP
1100 PEACHTREE STREET
SUITE 2800
ATLANTA, GA 30309 (US)

(21) Appl. No.: **10/057,849**

(22) Filed: **Jan. 24, 2002**

Related U.S. Application Data

(63) Continuation-in-part of application No. 09/633,410, filed on Aug. 7, 2000, which is a continuation-in-part of application No. 09/578,011, filed on May 24, 2000, which is a continuation-in-part of application No. 09/568,301, filed on May 9, 2000, now Pat. No. 6,427,141, which is a continuation of application No. 09/303,387, filed on May 1, 1999, now Pat. No. 6,128,608.

(60) Provisional application No. 60/263,696, filed on Jan. 24, 2001. Provisional application No. 60/298,757, filed on Jun. 15, 2001. Provisional application No. 60/275,760, filed on Mar. 14, 2001. Provisional application No. 60/083,961, filed on May 1, 1998.

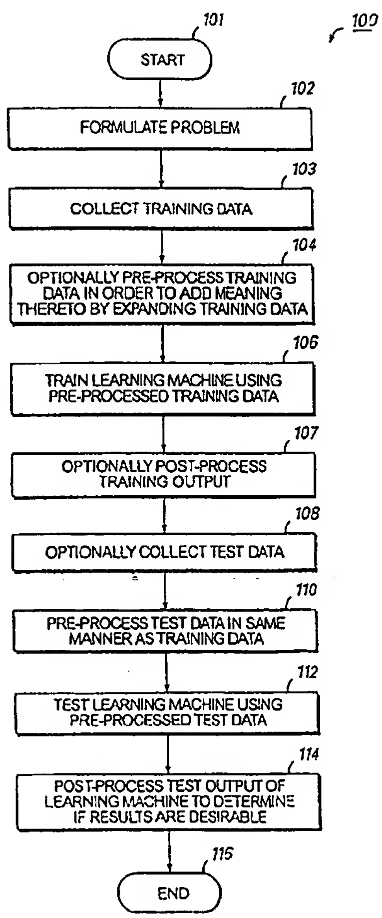
Publication Classification

(51) Int. Cl.[7] **G06N 5/02**

(52) U.S. Cl. **706/48**

(57) **ABSTRACT**

The methods, systems and devices of the present invention comprise use of Support Vector Machines and RFE (Recursive Feature Elimination) for the identification of patterns that are useful for medical diagnosis, prognosis and treatment. SVM-RFE can be used with varied data sets.



PGPUB-DOCUMENT-NUMBER: 20030172043

PGPUB-FILING-TYPE: new

DOCUMENT-IDENTIFIER: US 20030172043 A1

TITLE: Methods of identifying patterns in biological systems and uses thereof

PUBLICATION-DATE: September 11, 2003

INVENTOR-INFORMATION:

NAME	CITY	STATE	COUNTRY
RULE-47			
Guyon, Isabelle	Berkeley	CA	US
Weston, Jason	St. Leonard's on Sea		GB

US-CL-CURRENT: 706/48

ABSTRACT:

The methods, systems and devices of the present invention comprise use of Support Vector Machines and RFE (Recursive Feature Elimination) for the identification of patterns that are useful for medical diagnosis, prognosis and treatment. SVM-RFE can be used with varied data sets.

---------- KWIC ---------

Detail Description Paragraph - DETX (33):

[0094] As mentioned above, the exemplary optimal categorization method 300 may be used in pre-processing data and/or post-processing the output of a learning machine. For example, as a pre-processing transformation step, the exemplary optimal categorization method 300 may be used to extract classification information from **raw data**. As a post-processing technique, the exemplary optimal range categorization method may be used to determine the optimal cut-off values for markers objectively based on data, rather than relying on ad hoc approaches. As should be apparent, the exemplary optimal categorization method 300 has applications in pattern recognition, classification, regression problems, etc. The exemplary optimal categorization method 300 may also be used as a stand-alone categorization technique, independent from SVMs and other learning machines.

Detail Description Paragraph - DETX (177):

[0225] A more detailed discussion of the methods of a preferred embodiment follow. A SVM-RFE was run on the **raw data** to assess the validity of the method. The colon cancer data samples were split randomly into 31 examples for training and 31 examples for testing. The RFE method was run to progressively downsize the number of genes, each time dividing the number by 2. The preprocessing of the data for each gene expression value consisted of subtracting the mean from the value, then dividing the resultby the standard deviation.

Detail Description Paragraph - DETX (190):

[0238] The initial preprocessing steps of the data were described by Alon et al. The data was further preprocessed in order to reduce the skew in the data distribution. FIG. 13 shows the distributions of gene expression values across

tissue samples for two random genes (cumulative number of samples of a given expression value) which is compared with a uniform distribution. Each line represents a gene. FIGS. 13A and 13B show the **raw data**; FIGS. 13C and 13D are the same data after taking the log. By taking the log of the gene expression values the same curves result and the distribution is more uniform. This may be due to the fact that gene expression coefficients are often obtained by computing the ratio of two values. For instance, in a competitive hybridization scheme, DNA from two samples that are labeled differently are hybridized onto the array. One obtains at every point of the array two coefficients corresponding to the fluorescence of the two labels and reflecting the fraction of DNA of either sample that hybridized to the particular gene. Typically, the first initial preprocessing step that is taken is to take the ratio a/b of these two values. Though this initial preprocessing step is adequate, it may not be optimal when the two values are small. Other initial preprocessing steps include (a-b)/(a+b) and (log a-log b)/(log a+log b).

Detail Description Paragraph - DETX (192):
[0240] FIG. 14 shows the distribution of gene expression values across genes for all tissue samples. FIG. 14A shows the **raw data** and FIG. 14B shows the inv erf. The shape is roughly that of an erf function, indicating that the density follows approximately the Normal law. Indeed, passing the data through the inverse erf function yields almost straight parallel lines. Thus, it is reasonable to normalize the data by subtracting the mean. This preprocessing step is supported by the fact that there are variations in experimental conditions from microarray to microarray. Although standard deviation seems to remain fairly constant, the other preprocessing step selected was to divide the gene expression values by the standard deviation to obtain centered data of standardized variance.

Detail Description Paragraph - DETX (214):
[0261] **Unsupervised Clustering**

Detail Description Paragraph - DETX (215):
[0262] To overcome the problems of gene ranking alone, the data was preprocessed with an **unsupervised clustering** method. Genes were grouped according to resemblance (according to a given metric). Cluster centers were then used instead of genes themselves and processed by SVM-RFE to produce nested subsets of cluster centers. An optimum subset size can be chosen with the same cross-validation method used before.

Detail Description Paragraph - DETX (218):
[0265] With **unsupervised clustering,** a set of informative genes is defined, but there is no guarantee that the genes not retained do not carry information. When RFE was used on all QT.sub.clust clusters plus the remaining non-clustered genes (singleton clusters), the performance curves were quite similar, though the top set of gene clusters selected was completely different and included mostly singletons. The genes selected in Table 1 are organized in a structure: within a cluster, genes are redundant, across clusters they are complementary.

Detail Description Paragraph - DETX (227):
[0274] Compared to the **unsupervised clustering** method and results, the supervised clustering method, in this instance, does not provide better control over the number of examples per cluster. Therefore, this method is not as good as **unsupervised clustering** if the goal is the ability to select from a variety of genes in each cluster. However, supervised clustering may show specific clusters that have relevance for the specific knowledge being determined. In this particular embodiment, in particular, a very large cluster of genes was found that contained several muscle genes that may be related to tissue composition and may not be relevant to the cancer vs. normal separation.

Thus, those genes are good candidates for elimination from consideration as having little bearing on the diagnosis or prognosis for colon cancer.